Twitter rumour detection in the health domain

Citation for published version (APA):

Sicilia, R., Lo Giudice, S., Pei, Y., Pechenizkiy, M., & Soda, P. (2018). Twitter rumour detection in the health domain. *Expert Systems with Applications*, 110, 33-40. <https://doi.org/10.1016/j.eswa.2018.05.019>

Document license:
TAVERNE

DOI:
[10.1016/j.eswa.2018.05.019](https://doi.org/10.1016/j.eswa.2018.05.019)

Document status and date:
Published: 15/11/2018

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Twitter rumour detection in the health domain

Rosa Sicilia^{a,*}, Stella Lo Giudice^{a,1}, Yulong Pei^b, Mykola Pechenizkiy^b, Paolo Soda^a

^a Department of Engineering, University Campus Bio-Medico of Rome, Via Alvaro del Portillo 21, Roma 00128, Italy

^b Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, MB 5600, The Netherlands

ARTICLE INFO

Article history:

Received 21 December 2017

Revised 1 April 2018

Accepted 18 May 2018

Available online 24 May 2018

Keywords:

Health rumour detection

Social microblog

Twitter

Network-and user-based features

ABSTRACT

In the last years social networks have emerged as a critical mean for information spreading bringing along several advantages. At the same time, unverified and instrumentally relevant information statements in circulation, named as rumours, are becoming a potential threat to the society. For this reason, although the identification in social microblogs of which topic is a rumour has been studied in several works, there is the need to detect if a post is either a rumor or not. In this paper we cope with this last challenge presenting a novel rumour detection system that leverages on newly designed features, including influence potential and network characteristics measures. We tested our approach on a real dataset composed of health-related posts collected from Twitter microblog. We observe promising results, as the system is able to correctly detect about 90% of rumours, with acceptable levels of precision.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

Since their birth social networks have revealed themselves as a new and powerful mean for information and social behaviours spread. In the latest years they have gained a critical role in this scenario bringing along several advantages. In particular, social microblogs have emerged as the most used social network services, allowing users to be constantly connected and to stay abreast with ongoing events. Among all the positive aspects, there are also some downsides: as good and reliable information spreads, also rumours and false news easily reach a wide range of people in short time, becoming a potential threat to the society. For instance, in 2013 the official Twitter account of the Associated Press was hacked and it sent out a rumour about the explosion of two bombs at the White House and the US President being injured in the attack. This news caused deep panic in such a broad scale that ended with a dramatic, though brief, crash of the stock market (Domm, 2013). Along with this event, other similar episodes of misinformation causing disasters can be found in many domains, such as emergency situation management, health and wellness, politics, etc. (Chung & Zhang, 2017; Guan et al., 2014; Zubiaga, Liakata, Procter, Hoi, & Tolmie, 2016).

Unreliable and treacherous news is generally referred to as *rumour*. Although sometimes rumours are considered as constituting information which is ultimately deemed false (Zubiaga et al., 2016), this definition leaves out all those newsworthy stories that have the same level of uncertainty and potential threat at a first time spread and that only in the end reveal themselves to be true. The most used and comprehensive definition is the one that considers rumours as unverified and instrumentally relevant information statements in circulation (DiFonzo & Bordia, 2007).

Straightforwardly, the social microblogs available online could be affected by rumours. Twitter is a social microblog that counts millions of users from all over the world, facilitating real-time propagation of information to a large group of people and allowing users to carry out different actions, such as posting 140-character-long messages (tweets), replying to such posts (replies), or forwarding them (reposts or retweets). With these actions a user can start or participate a *conversation*, which therefore consists of a set of tweets, replies, and retweets.² Furthermore, a set of conversations can be grouped into a *topic* by keywords. Twitter is therefore an ideal environment for the diffusion of breaking-news directly from the news source and the geographical location of events. Online social microblogs can be obviously affected by rumours. The possibility of recognising them could be a powerful mean to prevent misinformation spreading, along with the possible consequences that this could cause.

* Corresponding author.

E-mail addresses: r.sicilia@unicampus.it (R. Sicilia), logiudice@2mel.nl (S. Lo Giudice), y.pei.1@tue.nl (Y. Pei), m.pechenizkiy@tue.nl (M. Pechenizkiy), p.soda@unicampus.it (P. Soda).

¹ Present address: 2M Engineering Ltd, John F. Kennedylaan 3, Valkenswaard, XC 5555, The Netherlands

² In the following, we will use the general term *post* to refer to a tweet, a reply or a retweet.

For these reasons rumour detection in social media is an issue in high demand that has been only partially investigated in the literature. On the one side, there exist three papers that identify if a topic is affected by rumours (Castillo, Mendoza, & Poblete, 2011; Kwon, Cha, Jung, Chen, & Wang, 2013; Ma, Gao, Wei, Lu, & Wong, 2015), but they do not determine which posts are rumours or not. On the other side, Wu, Yang, and Zhu (2015) presented an approach for rumour detection at conversation level. Hence, to the best of our knowledge, rumour detection within a specific topic and within a conversation is still a point deserving further research efforts, since, in many cases, there is the need to identify which posts are reliable on a desired argument. In this respect, this paper aims at filling this gap and it presents a novel rumour detection system for Twitter focusing on topic-specific rumours, namely the health-related topic. Indeed, among all the information domains, this field is of particular interest for two main reasons: first, the literature has shown that rumours concentrate in some specific topics (Wu et al., 2015) such as the health, and second, nowadays people often look for health knowledge and advices on online services, but not all these resources provide accurate or reliable information (Boyer, Baujard, & Geissbühler, 2011). Hence, it is intuitive why discovering rumours at a smallest granularity could be of great value in this field. We explore two levels of features, named as user level and network level, respectively. For each level, we study three types of features including influence potential, personal interest and network characteristics. In particular, besides widely used features such as user statistics and sentiment of tweets, we also develop new features including the likelihood that a tweet is retweeted and the likelihood of a URL to be shared, conversation size, fraction of users followers of root, and fraction of tweets with URLs in a conversation. The use of general features unrelated to the particular topic considered would allow to develop an approach that could be employed in the future in other analyses on different specific domains. We also investigate the discrimination power of different features using different classifiers.

To sum up, our main contributions are: (1) we develop a novel health-related rumour detection system on Twitter using several new features; (2) our proposal works at level of single post rather than at topic level; (3) we explore the selection of different types of features using different classification methods for rumour detection. (4) we validate this system as well as these features on a real Twitter dataset, being able to correctly detect about 90% of rumours;

The paper is organised as follows: next section presents the background and motivations, Section 3 describes the dataset, Section 4 introduces the methods dealing also with Twitter structure, feature extraction and selection. Section 5 reports the experimental results, whereas Section 6 provides concluding remarks.

2. Background and motivations

Traditionally, it has been very difficult to study people's reactions to uncertain news mainly because of the complex problem of real-time collection of reactions as the rumour unfolds (Zubiaga et al., 2016). Nevertheless, with the widespread adoption of Internet and online social media, the availability of a massive quantity of data and real-time information allowed the use of computational methods based on machine learning techniques in order to improve rumour identification.

Castillo et al. (2011) presented a method detecting if a Twitter topic (i.e., a collection of posts and reposts) is either a rumour or not. To this aim, and inspired by previous work (Agichtein, Castillo, Donato, Gionis, & Mishne, 2008; Alonso, Carson, Gester, Ji, & Nabar, 2010; Hughes & Palen, 2009), they used a set of message-based, user-based, topic-based and propagation-based features, which have also been adopted in later studies (Ma et al.,

2015; Wu et al., 2015; Yang, Liu, Yu, & Yang, 2012). These features fed a J48 decision tree to distinguish between credible and rumour tweets. Their approach was tested using a dataset collecting Twitter newsworthy topics between April and September 2010, which belong to the news and to the conversation groups (Java, Song, Finin, & Tseng, 2007; Pear Analytics, 2009), such as earth, notebook, movie, bed, hangover, woke. Using different feature subsets, experiments adopting a three-fold cross validation showed that precision and recall ranged from 70% up to 80%.

Wu et al. (2015) dealt with rumour detection at conversation level in the social microblog Sina Weibo (Wu et al., 2015). They introduced new features related to the user, to the message and to the repost in addition to those describing the basic characteristics of the message proposed in (Agichtein et al., 2008; Qazvinian, Rosengren, Radev, & Mei, 2011; Yang et al., 2012).³ In particular, repost features catch temporal information as they take into account the temporal distance between the original posts and the reposts. The authors proposed an hybrid Support Vector Machine (SVM), combining a traditional radial basis function with a random-walk graph kernel modelling the propagation patterns of the messages in a tree. The structure of such a tree assumes that network users can be divided into opinion leaders (e.g., influential users) and normal users, allowing to highlight some propagation patterns in rumour diffusion, and pointing out that the influence of multiple opinion leaders can quickly create a hype followed by ordinary users. Experiments were conducted adopting a three-fold cross validation using a dataset of the Sina Weibo community management centre, which includes 2601 rumours and 2536 normal messages, with 4 million distinct users involved in these messages, on many different topics collected between 2012/05/28 and 2014/04/11, attaining precision and recall larger than 90%.

Whilst the approaches described so far mainly exploited the propagation structure of the network, Kwon et al. (2013) and Ma et al. (2015) explored the temporal characteristics of the rumour spreading life-cycle. In particular, Kwon et al. (2013) examined temporal, structural and linguistic aspects of rumour diffusion on Twitter. They proposed a periodic time series model considering daily and external shock cycles and demonstrating that rumour likely have fluctuations over time. Using features extracted from the temporal, the structural and the linguistic diffusion patterns of the tweets, they trained different classification algorithms (such as Random Forest, SVMs, etc.). This approach was tested on a dataset of 130 different topics (such as Big foot, Harvard, Summize, Catfish, etc.), out of which 68 were rumours and 57 were non-rumours, yielding to a best accuracy of 90%. In the work by Ma et al. (2015), the main idea is to model the variation of social context features during the message propagation. The authors defined a time stamp for each microblog taking into account the moment the microblog was posted and the time interval between the first and the last post published about the same event. They proposed to capture the variation of the features for each time stamp, considering the slopes of the features between two consecutive intervals. The approach was tested on Twitter and Sina Weibo general purpose datasets used by Castillo et al. (2011) and Wu et al. (2015), and it detected rumour topics in the early stage of their diffusion with an accuracy ranging from 80% up to 90%.

Whatever the rumour detection approach, the review of the literature reported so far shows that existing proposals detect if a topic or a conversation contains rumour, while they do not investigate which post is a rumour or not. In particular, they used the

³ It is worth noting that Qazvinian et al. (2011) and Yang et al. (2012) did not deal with rumour detection on social media posts, but they presented a system tailored to identify false and true positive rumours already detected. The former dealt with data listed in About.com Urban Legends, while the latter used the Sina Weibo official rumour busting service.

knowledge on the tweet topic as a feature in the detection process (Castillo et al., 2011; Ma et al., 2015; Wu et al., 2015). Next, most of the contributions consider only the most influential users or the most crowded conversations in the network, deleting all the small conversations (Castillo et al., 2011; Ma et al., 2015; Wu et al., 2015). This helps the rumour detection process in that the larger the network, the more the rumours that can spread. However, when scaling the systems in the smaller context of a single topic, small conversations cannot be removed as they can carry useful information, as the beginning of a rumour spreads from a normal user. Moreover, since semantics of words and sentences may differ as topic changes, strategies based on descriptors of such characteristics as well as the use of knowledge on the tweet topic may be not directly applicable when the rumours detection system has to deal with a specific topic. For all these reasons, the systems available in the literature can not be directly applicable to detect rumour at level of each post.

3. Dataset

This work copes with rumour detection in each post of health-related news, a specific topic domain that, to the best of our knowledge, has not been investigated yet. For this reason a public dataset was not available and we therefore queried Twitter using the keywords *#zika* and *zika microcephaly*, which was one of the main sanitary trends in 2016. Indeed, Zika virus disease was declared a Public Health Emergency of International Concern (PHEIC)⁴ by the World Health Organization on February 1st, 2016. We retrieved 497 posts in two days of April 2016, and 1085 posts in three days of May 2016 with a gap between the two time intervals. Even if such time intervals could seem short, they ensure that the spreading of the news throughout the network is not the mere outcome of homophily, but the result of a real interest of the users in that specific news (Dave, Bhatt, & Varma, 2011). Moreover, the gap between the first and the second acquisition ensures that the dataset contains different rumour tweets: indeed, if we considered only tweets collected in consecutive days, the dataset could contain many repetitions of the same post as rumours show a bursting behaviour tending to appear and to fade repeatedly in a few days (Kwon & Cha, 2014; Kwon et al., 2013). The infectious nature of the rumour suggested us to consider not informative, and therefore to discard, the tweets that did not generate any retweet or reply and the tweets whose propagation graphs could not be reconstructed. Finally, the dataset contains 709 samples that were manually labelled by two human annotators into three classes, namely *rumour* (~54% of samples), *non-rumour* (~30% of samples) and *unknown* (~16% of samples). Each tweet was blindly annotated and possible disagreements were solved through debate for a common understanding. The annotators were also made aware of the scope of the work, so they knew the following definitions of the three classes (DiFonzo & Bordia, 2007):

- **Rumour:** as presented in Section 1, a rumour is an unverified and instrumentally relevant information statement in circulation. It indicates news without reference, so that can't be verified, and that is around. Typically it shows an instantaneous and huge spread in the network, in other words an *infectious behaviour*. An example retrieved from the dataset is:

RT @ClassicPict: Mosquitoes kill more annually than Sharks.
#ZikaVirus <https://t.co/P2xi13Vx2Z>.

- **Non-rumour:** this class identifies all referenced news with at least one link to a certified and official webpage, such as news

papers, hospitals, universities, etc. Below there is an example of a tweet linked to the WHO webpage.

RT @WHO is dispelling rumours around Zika & microcephaly.
See the facts about #ZikaVirus: <http://goo.gl/JDKuys>.

- **Unknown:** this class collects all the news which cannot be determined to be false, such as news with a link to an empty page, news potentially true but without reference, news with a main topic unrelated to the one of interest. An example of potentially true news without any reference is:

RT @DoughertyNews9: ICYMI: Researchers discover #ZikaVirus can be carried by a 2nd species of mosquitoes. We have details soon on @NEWS9

The dataset was divided into training (90% of samples) and validation (10% of samples). Assuming that classification tests are described as Bernoulli-processes this partition ensures that, when estimating the performance, the width of the 95% confidence interval is equal to 0.1, which is a reasonable choice.

4. Methods

We present a rumour detection system whose pipeline is shown in Fig. 1. The upper panel depicts the training phase: first the system acquires the data from Twitter, then it identifies two structural levels (named as *user level* and *network level* in the following) that lead to the extraction of representative features. Feature evaluation and feature selection on a disjoint validation set permit to build the learning model, which is finally trained. To perform the classification on a new set of tweets, the lower panel of Fig. 1 shows that the system exploits both the best feature set and the learning model defined in the training stage.

After a preliminary focus on Twitter structure offered in the next subsection, we describe feature extraction and evaluation.

4.1. Twitter structure

Fig. 2 shows a representation of the Twitter structure, which can be divided into user and network levels.

The former identifies the characteristics of the user and his/her statuses, which are considered as self-standing and without links in the network. For instance, some properties observable at this level are the number of followers and followings of a single user or the sentiment related to a specific post.

The latter identifies the interaction between users in the network and the related properties given by retweets and replies. Such two actions can be represented by conversation graphs, where each node denotes a user and each link between the nodes corresponds either to a retweet or to a reply. Fig. 2 shows two examples: on the left there is a retweet graph with a star shape, where each action connects a retweet (red node) to its original post (grey star). On the right there is a graph of replies, which has a more intricate structure with an original post leading to multiple intermediate points (blue nodes) and multiple end points as well (green nodes). The characteristics of the conversation graphs allow to analyse the spread of information in the network and the influence properties of the users.

4.2. Feature extraction

We now present the set of features, which has been divided into three different groups: referred to as *influence potential*, *personal interest* and *network characteristics* (Fig. 1). Each different group catches a different property of the social microblog, and within each group measures related to both users and to the network levels can be clearly distinguished. The interested readers can

⁴ About Zika Virus Disease: <http://www.cdc.gov/zika/about/>

Fig. 1. Pipeline of the rumour detection process.

Fig. 2. Representation of the Twitter structure. The retweet conversation is represented by a star graph with all retweets (red nodes) referring to an original post (grey node). Conversely, the replies show a more complex structure, with an original post that leads to multiple intermediate points (blue nodes) and to multiple end points as well (green nodes). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

refer to the supplementary material for the detailed feature definition.

4.2.1. Influence potential measures

These measures describe the power or the capacity of causing an effect in indirect or intangible ways.

Some user and network level features were introduced by Castillo et al. (2011) and by Wu et al. (2015), where the authors suggested to characterize a user considering all the properties related to his/her Twitter account. Therefore, we computed the number of followers and followings of the user, as well as if he/she was follower of another user involved in a conversation, and the age of his/her account, defined as the difference in months between the registration date and the current date. We considered also features related to the specific tweet (presence of URLs or question marks, etc.). Nevertheless, we also introduced two new user level features named as P_{rt} and P_{url} . P_{rt} is the probability that a tweet is retweeted, computed as the number of times the tweet is shared (if it is a retweet) divided by the total number of samples in the dataset. P_{url} is the probability of a URL to be shared; it counts how

Table 1

List of all the features extracted, where “#” and “Avg” stand for “number of” and “average”, respectively. Newly introduced features are marked with a dagger (†). The features inspired by graph theory that, to the best of our knowledge, have not been used in rumour detection yet are marked with a diamond (◊).

Feature group	User level	Network level
Influence potential	#Followers	
	#Followings	
	#Statuses	Avg #Followers
	Registration Age	
	is a Retweet? (isRT)	Avg # Followings
	has URL? (hasURL)	
	has “?”	
	P_{rt} †	Avg # Statuses
	P_{url} †	
	is Follower?	Avg Registration Age
Personal interest	Sentiment score	Avg Sentiment score
		Fraction of positive
		Fraction of negative
Network characteristics		PageRank◊
		Closeness centrality◊
		Betweenness centrality◊
		Conversation size†
		Fraction of Users followers of the root (FUF)†
		Fraction of tweets with URL (FTWU)†

many times the attached URL was used as a reference source divided by the total number of URLs in the dataset.

4.2.2. Personal interest measures

Understanding the reaction of people to a specified news in terms of opinions and related sentiments is a fundamental step for rumour detection: in fact they convey much information about what people actually think of a specific news, if they believe it or not. To extract this information from the retrieved tweets we performed a lexicon-based sentiment analysis extracting the SentiWordNet score (Esuli & Sebastiani, 2006), including a stage of negation detection to take into account the presence of negated contexts. For example, if a status says “I don’t like this post”, the score of the word *like* switches from a positive to a negative value, because of the presence of the negation *don’t*. The sentiment score is computed summing up the scores for every match of each token in the tweet with a word belonging to lexicon synset, considering that all the negation words are changed in sign. While the senti-

ment score is directly used as user level feature, network level features were computed considering the scores assigned to all tweets belonging to each conversation (Castillo et al., 2011).

4.2.3. Network characteristics measures

These features were designed to describe the characteristics of the propagation graphs built with retweets and replies conversations. To identify how the information spreads in the network we quantified the degrees of influence of the users. To this aim, we considered measures of centrality and popularity originally conceived for the graph theory, such as the PageRank, an index of the popularity of the node in a network, the closeness centrality, a measure of the independence or efficiency of a node, and the betweenness centrality, a representation of the potential of a point for control of communication (Freeman, 1978). We also introduced other three structural measures: (i) the conversation size, which is the number of nodes in a conversation, (ii) the fraction of followers of the root, given by the number of users in a conversation who are followers of the root user, divided by the conversation size, (iii) the fraction of tweets with an URL, computed as the number of tweets with a URL in a conversation divided by the conversation size.

4.3. Feature evaluation and selection

The evaluation of the designed measures is an important step in the process of feature engineering: it aims at analysing which features are the most informative and whether the newly introduced ones are meaningful or not for the classification purpose.

Among the different techniques for feature evaluation (Huang, 2015), we applied the wrapper method that compares different candidate sets of variables using the performances of classification algorithms. We analysed every feature by evaluating how the Area Under the ROC curve (AUC) changes when it is removed from the feature set. We used the AUC since it is independent to the decision threshold, and it is invariant to the a priori class probability distribution. The discriminant power of a feature is measured through the following score:

$$S(f) = AUC_{complete\ set} - AUC_{leave\ f\ out\ set} \quad (1)$$

where f is the considered descriptor in the feature set. Hence, the larger the value of $S(f)$, the more discriminant the feature f .

To avoid any bias given by the specific classification algorithm used, we ran such an evaluation using several classifiers belonging to different learning paradigms (listed in Section 5). Feature evaluation was conducted in 10-fold cross validation on the training set and the results were then summarized via a rank analysis computing the relative performances of one feature with respect to the others. First, for each classifier used, we sorted individually the values of $S(f)$ and then we assigned to each variable f a rank with respect to its place among the others. The highest rank is 24 (assigned to the worst case corresponding to the least discriminant feature) and the lowest is 1 (assigned to the best case). Ranks of each feature are finally summed up per classifier, and then they are normalized with respect to the highest possible value (i.e. highest rank $\times$ numbers of classifiers).

The feature selection step aims at finding the most representative measures among all those computed. To this end, we first applied a threshold on the aforementioned ranks returning a set of features. Second, on such a reduced feature space, a classifier is trained on the training set. Third, this classifier labels samples belonging to an independent validation set, where we measured the AUC. Such a three-steps procedure is repeated for different values of the threshold so that, at the last iteration, the feature set is composed only by one descriptor. The selection is then repeated

for different learning paradigms and we find out the reduced feature set providing the largest average AUC per classifier. Next, we discuss the experimental results.

5. Experimental results

We considered classifiers belonging to different learning paradigms, including a Multi-Layer Perceptron (MLP) as a neural network, a Nearest Neighbour (NN) as a statistical classifier, a Support Vector Machine (SVM) as a kernel machine, a Random Tree (RT) as a decision tree, a multiclass Adaboost (MAda) as a multi-expert system and a Random Forest (RF) as an ensemble of trees. We set their parameters to the default values. Although we acknowledge that their tuning could lead to better results, we preferred to maintain a baseline configuration as the basis for comparison between them. Indeed, it is reasonable that the classifier which wins on average on all the experiments would also win if a better setting was performed (Fernández, López, Galar, Del Jesus, & Herrera, 2013). Furthermore in a framework where the classifiers are not tuned, the winning learning model tends to correspond to the most robust one, which is also a desirable characteristic.

Next subsections describe the experimental results of the feature evaluation stage and of the feature selection phase, and finally Section 5.4 reports the performances achieved in the testing scenario.

5.1. Feature evaluation

Fig. 3 shows the ranks of each feature f computed as described in Section 4.3. Each bar in the plot is the normalized sum of the ranks provided by different learning paradigms, whose contributions are represented in different colours. Note that the shorter the bar the more informative the feature is.

The newly introduced features (marked with a dagger) show considerable descriptive power compared to the others as their ranks are usually lower than the median. In particular the user level measures P_{rt} and P_{url} as well as the network level feature FTWU are worthy of mention. The P_{rt} value conveys information about how much “popular” a node is, computing its probability of being retweeted. Since, usually, the rumours present a huge spread in the network, these tweets have a high value of the P_{rt} , thus helping the distinction between rumours and non-rumours. P_{url} represents the probability of a URL to be shared in a post: its informative value lies in the strong habit of users to support a statement linking a web page to the post. Therefore, if this score is high, it means that the considered tweet is probably a non-rumour, and vice versa. The FTWU, measuring the fraction of tweets with a URL attached, appears particularly informative for the problem. Indeed, people manifest their disagreement with a not reliable news using some different referenced sources.

Among the newly introduced features, the closeness and betweenness centrality measures attain high rank values (Fig. 3), suggesting that they are not so informative. By definition, such features discriminate less when computed on very small networks (e.g. a two node network), which frequently occur in our dataset. This happens since we adopted a general approach where small conversations and not influential users were not filtered out, as reported in Section 2.

We also note that the sentiment score attains a low rank. This finding agrees with data reported in the literature (Castillo et al., 2011; Qazvinian et al., 2011; Wu et al., 2015; Yang et al., 2012), showing that the evaluation of both the sentiment in single tweets and the average sentiment in a conversation is particularly important for the rumour detection problem. That's because the information related to the opinions can be highly discriminative, especially if we consider the distinction between neutral posts and

Fig. 3. Stacked histogram of the rank analysis, where the shorter the bar the more informative the feature is. The dotted red line is the median rank value. The bars also show the contribution of each classification algorithm to the importance of the single variable, highlighted with different colours. The newly introduced features are marked with a dagger (†), while features inspired by graph theory are marked with a diamond (◊). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tweets with high scores (the absolute value of both negative and positive sentiments). In fact, it's more likely that a neutral status simply reports news with its reference, whereas emphatic statuses try to gain the attention of users in the social network, and use the infection to spread unreliable news.

The analysis of the results also shows that the most informative features are actually those belonging to the network level since almost all of them have a normalised rank lower than the median. Indeed, except for P_{rt} , P_{url} and for the sentiment score, we found that the user level attributes (which related to the single nodes of the network) do not discriminate well the samples in the feature space. Although this result disagrees with previous findings reported in the literature, we deem this happens because we do not delete the small conversation, as most of the previous work do (Castillo et al., 2011; Ma et al., 2015; Wu et al., 2015). In fact, such work considered only those networks where the news spread was strongly dependent on the nature of the nodes (e.g. with a high level of opinion leadership (Wu et al., 2015)). In this case the features related to the nodes represented valuable information for the rumour detection task. Consequently the networks originating from a node with a low level of opinion leadership were all filtered out. Unlike what reported in the literature, the presence of small conversations in our dataset allows us to maintain the classification problem as much general as possible, so that the classifier cannot discriminate the spread of a rumour only from the characteristics of the single agents of the networks.

5.2. Feature selection

The feature selection experiments were conducted as described in Section 4.3 on the basis of the results showed in Fig. 3.

We estimated the performances of each classifier on the validation set and then we averaged the results among all the classifiers. The results show that the best feature subset is composed of 20 features, discarding the *number of followers*, the *number of statuses*, *has "??"* and *is Follower*, which also showed the highest ranks (Fig. 3). This finding agrees with the observations presented in Section 5.1: the measures related to the user's characteristics have been discarded, in that they provide little information.

5.3. Classifier performances

In this section we look for the best learning paradigm for our classification problem, whereas next subsection reports the system validation on an independent dataset.

To determine the best classifier we compared the aforementioned learning paradigms running a 20-fold cross validation on the training set, representing the samples in the feature space selected as described in Section 5.2. As performances indexes we measured the overall accuracy and the average AUC per class.⁵ Table 2 shows the results of such an exhaustive comparison: in each panel the lower triangle of the tabular shows the amount of win-tie-loss of a classifier in a row comparing with a classifier in a column, whereas the upper triangle reports the *p*-value of the pairwise difference between the classifier performances measured using the Wilcoxon rank sum test. In panel *a* it is straightforward to observe that the RF significantly outperforms most of the other learning paradigms: indeed, the *p*-value is many times smaller than 0.01 and the number of wins is close to 20, i.e. the maximum value. Similar observations hold also in case of AUC (panel *b* of the table), where in particular RF wins over all the other classifiers. Note also that RF, as well as SVM, has been the most used in the literature on rumour detection (Castillo et al., 2011; Ma et al., 2015; Wu et al., 2015; Yang et al., 2012).

5.4. System validation

We also validated our rumour detection system computing the performances on a test set using the best feature subset and the RF classifier, as suggested by the aforementioned results. The test set was collected using the same procedure described in Section 3, and it consists of 91 labelled samples, where the a-priori probabilities of rumour, non-rumour and unknown are equal to 55%, 31% and 14%, respectively. It is worth observing that such a set contains a number of samples almost equal to the validation set introduced in Section 3, thus ensuring that the confidence interval

⁵ A performances metric is named "per class" when it is computed considering the binary classification task derived from the decomposition of the original recognition task.

Table 2

Exhaustive comparison between the performances of different classifiers expressed in terms of accuracy (panel a) and AUC (panel b). In the lower triangle each tabular shows the amount of win-tie-loss of a classifier in a row comparing with a classifier in a column. The upper triangle reports the *p*-value of the pairwise difference between the classifier performances measured using the Wilcoxon rank sum test. Last column of the table reports the average values of accuracy and AUC for each classifier computed running a 20-fold cross validation on the training set.

	SVM	MLP	NN	MAda	RT	RF	Avg acc (%)
SVM	–	9.8E–03	1.8E–04	8.2E–01	8.4E–06	4.8E–06	71.96
MLP	15–3–2	–	1.6E–01	1.8E–02	1.8E–02	3.4E–03	78.59
NN	17–0–3	13–3–4	–	4.5E–04	2.9E–01	6.2E–02	81.58
MAda	3–16–1	2–4–14	3–1–16	–	2.3E–05	9.2E–06	72.42
RT	19–1–0	15–3–2	12–3–5	19–1–0	–	3.2E–01	84.09
RF	19–0–1	18–1–1	14–4–2	19–0–1	10–7–3	–	86.13
(a) Win-tie-loss and <i>p</i> -value comparison in terms of accuracy (acc)							
	SVM	MLP	NN	MAda	RT	RF	Avg AUC (%)
SVM	–	3.4E–07	6.6E–06	5.2E–01	2.3E–06	6.7E–08	73.63
MLP	20–0–0	–	7.9E–01	1.2E–06	9.1E–02	2.5E–04	89.06
NN	18–0–2	10–0–10	–	2.3E–05	1.7E–01	8.3E–05	87.86
MAda	14–0–6	0–0–20	3–0–17	–	1.2E–05	6.8E–08	74.92
RT	19–0–1	6–0–14	7–0–13	19–0–1	–	3.5E–06	85.66
RF	20–0–0	20–0–0	19–0–1	20–0–0	20–0–0	–	95.65
(b) Win-tie-loss and <i>p</i> -value comparison in terms of AUC							

of estimated performance is the same. We obtained the following performances: the overall accuracy is equal to 73.63%, the average precision per class is 72.80%, the average recall per class is 73.60% and the average AUC per class is 89.00%. Furthermore, as we are mainly interested in the recognition of rumour samples, which we regarded as “positive” samples, the true positive rate and the false positive rate achieved on the rumour class are equal to 92.00% and 39.00%, respectively. This means that such a system is *liberal*, i.e., it tends to classify samples as rumour, giving sometimes false alarms that could be tolerated for the purposes of this work. Indeed, in the health tweet domain it is reasonable that the cost of a false positive is lower than the one of false negative. This suggests that the use of a liberal system should be advocated in all those delicate cases where the spread of an unreliable information is far more dangerous than questioning the trustworthiness of official statements.

Finally, it is worth noting that the performances reported in this work cannot be compared with those presented in the literature since the recognition task defined here is different from those previously defined by Castillo et al. (2011), Kwon et al. (2013), Ma et al. (2015), and Wu et al. (2015). Indeed, they do not explicitly classify the single post retrieved since they work at topic or conversation level, as already discussed in Section 1 and in the last paragraph of Section 2. Moreover, both Castillo et al. (2011) and Wu et al. (2015) make use of features that cannot be computed from our dataset. In the former case, several of its features are measures computed from social media posts aggregated by topic. In the latter contribution, some of the descriptors are designed for SinaWeibo social microblog thus leveraging on information that cannot be computed from Twitter data; furthermore, other features work only when different topics are available. Furthermore, Ma et al. (2015) and Kwon et al. (2013) took into account the temporal characteristics of the propagation, resorting to time series models which cannot be applied on our dataset that actually lacks of the needed temporal proximity.

6. Conclusions

In this work we have presented an approach to detect rumour in each post of a single topic domain related to health news using Twitter data. The restricted scale made the detection problem different from the literature, as it excludes the possibility of using the topic information as a feature, of making any a priori as-

sumption on the nature of the network, and of making a priori assumptions on user's characteristics. To represent our samples we used not only features available in the literature, but we also introduced new descriptors inspired by the study of the graph theory and the social influence models including the likelihood that a tweet is retweeted and the likelihood of a URL to be shared, conversation size, fraction of users followers of root, and fraction of tweets with URLs. Experimental study on a real dataset demonstrated the effectiveness of this system and new features which can correctly detect about 90% of rumours, with acceptable levels of precision. Moreover, we explored the selection of different features using different classification methods for rumour detection which is beneficial to the future studies in selecting effective combination of classifiers and features in such task.

Future work could be directed towards the application of this system in a different test case, considering a different topic (e.g. security and fraud detection) and a much larger dataset. Indeed, this is possible as we have proposed here features that are unrelated to the particular topic considered.

Data availability

The dataset generated during the current study can be made available from the corresponding author on reasonable request.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.eswa.2018.05.019](https://doi.org/10.1016/j.eswa.2018.05.019)

References

- Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 183–194). ACM.
- Alonso, O., Carson, C., Gerster, D., Ji, X., & Nabar, S. U. (2010). Detecting uninteresting content in text streams. *SIGIR Crowdsourcing for Search Evaluation Workshop*.
- Boyer, C., Baujard, V., & Geissbühler, A. (2011). Evolution of health web certification through the HONcode experience. In *Mie* (pp. 53–57).
- Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of the 20th international conference on World Wide Web* (pp. 675–684). ACM.
- Chung, S. S., & Zhang, S. (2017). Volatility estimation using support vector machine: Applications to major foreign exchange rates. *Electronic Journal of Applied Statistical Analysis*, 10(2), 499–511.
- Dave, K. S., Bhatt, R., & Varma, V. (2011). Modelling action cascades in social networks. *ICWSM*.

- DiFonzo, N., & Bordia, P. (2007). *Rumor psychology: Social and organizational approaches*. American Psychological Association.
- Domm, P. (2013). False rumor of explosion at white house causes stocks to briefly plunge; AP confirms its twitter feed was hacked. *CNBC.COM*, 23.
- Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*: 6 (pp. 417–422). Citeseer.
- Fernández, A., López, V., Galar, M., Del Jesus, M. J., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, 42, 97–110.
- Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3), 215–239.
- Guan, W., Gao, H., Yang, M., Li, Y., Ma, H., Qian, W., et al. (2014). Analyzing user behavior of the micro-blogging website sina weibo during hot social events. *Physica A: Statistical Mechanics and its Applications*, 395, 340–351.
- Huang, S. H. (2015). Supervised feature selection: A tutorial. *Artificial Intelligence Research*, 4(2), 22.
- Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3–4), 248–260.
- Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we Twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis* (pp. 56–65). ACM.
- Kwon, S., & Cha, M. (2014). Modeling Bursty Temporal Pattern of Rumors.. *ICWSM*.
- Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *Data Mining (ICDM), 2013 IEEE 13th international conference on* (pp. 1103–1108). IEEE.
- Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 1751–1754). ACM.
- Pear Analytics (2009). *Twitter study*. <http://web.archive.org/web/20110715062407/www.pearanalytics.com/blog/wp-content/uploads2010/05/Twitter-Study-August-2009.pdf>.
- Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: identifying misinformation in microblogs. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1589–1599). Association for Computational Linguistics.
- Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on Sina Weibo by propagation structures. In *IEEE 31st international conference on data engineering* (pp. 651–662). IEEE.
- Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on Sina Weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics* (p. 13). ACM.
- Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS One*, 11(3), e0150989.